CLAIMS:

What is claimed is:

1. A method of allocating resources of a computing system to hosting of a data network site to thereby maximize generated profit, comprising:

calculating a total profit for processing requests received by the computing system for the data network site based on at least one service level agreement; and

allocating resources of the computing system to maximize the total profit.

2. The method of claim 1, wherein calculating a total profit includes, for each request received by the computing system for the data network site, determining whether processing of the request generates a profit or a penalty, wherein a profit is generated when the allocation of resources is such that the request is processed in accordance with the service level agreement and a penalty is generated when the allocation of resources is such that the request is not processed in accordance with the service level agreement.

3. The method of claim 1, wherein calculating a total profit includes using a cost model in which profit is gained for each request to the data network site that is processed in accordance with a service level agreement and a penalty is paid for each request to the data network site that is not processed in accordance with the service level agreement.

4. The method of claim 1, wherein the requests are classified into one or more classes of requests and each class of request has a corresponding service level agreement from the at least one service level agreement.

5. The method of claim 1, wherein allocating resources includes determining an optimal traffic assignment for routing requests to thereby maximize the total profit.

6. The method of claim 1, wherein the computing system is a web server farm and wherein the resources are servers of the web server farm.

7. The method of claim 6, further comprising determining an optimum resource allocation to maximize the total profit.

8. The method of claim 7, wherein determining an optimum resource allocation includes:
modeling the resource allocation as a queuing network;
decomposing the queuing network into separate queuing systems; and
summing cost calculations for each of the separate queuing systems.

9. The method of claim 8, further comprising optimizing the summed cost calculations to maximize generated profit and thereby determine an optimum resource allocation.

10. The method of claim 1, wherein allocating resources includes determining an optimum traffic assignment and an optimum generalized processor sharing coefficient for a class of requests.

11. The method of claim 1, wherein allocating resources includes optimizing a cost function associated with a class of requests.

12. The method of claim 11, wherein optimizing the cost function includes modeling the optimization as a network flow from a source, through sinks representing sites/classes of request and servers/classes of requests, to a supersink.

13. The method of claim 8, wherein decomposing the queuing network into separate queuing systems includes decomposing the queuing network into decomposed models for each class in a hierarchical manner.

14. The method of claim 13, wherein a decomposed model for class k is based on a decomposed model of classes 1 through k-1.

15. An apparatus for allocating resources of a computing system to hosting of a data network site to thereby maximize generated profit, comprising:

means for calculating a total profit for processing requests received by the computing system for the data network site based on at least one service level agreement; and

means for allocating resources of the computing system to maximize the total profit.

16. The apparatus of claim 15, wherein the means for calculating a total profit includes means for determining whether processing of each request generates a profit or a penalty for each request received by the computing system for the data network site, wherein a profit is generated when the allocation of resources is such that the request is processed in accordance with the service level agreement and a penalty is generated when the allocation of resources is such that the request is not processed in accordance with the service level agreement.

17. The apparatus of claim 15, wherein the means for calculating a total profit includes means for using a cost model in which profit is gained for each request to the data network site that is processed in accordance with a service level agreement and a penalty is paid for each request to the data network site that is not processed in accordance with the service level agreement.

18. The apparatus of claim 15, wherein the requests are classified into one or more classes of requests and each class of request has a corresponding service level agreement from the at least one service level agreement.

19. The apparatus of claim 15, wherein the means for allocating resources includes means for determining an optimal traffic assignment for routing requests to thereby maximize the total profit.

20. The apparatus of claim 15, wherein the computing system is a web server farm and wherein the resources are servers of the web server farm.

21. The apparatus of claim 20, further comprising means for determining an optimum resource allocation to maximize the total profit.

22. The apparatus of claim 21, wherein the means for determining an optimum resource allocation includes:

means for modeling the resource allocation as a queuing network;

means for decomposing the queuing network into separate queuing systems; and

means for summing cost calculations for each of the separate queuing systems.

23. The apparatus of claim 22, further comprising means for optimizing the summed cost calculations to maximize generated profit and thereby determine an optimum resource allocation.

24. The apparatus of claim 15, wherein the means for allocating resources includes means for determining an optimum traffic assignment and an optimum generalized processor sharing coefficient for a class of requests.

25. The apparatus of claim 15, wherein the means for allocating resources includes means for optimizing a cost function associated with a class of requests.

26. The apparatus of claim 25, wherein the means for optimizing the cost function includes means for modeling the optimization as a network flow from a source, through sinks representing sites/classes of request and servers/classes of requests, to a supersink.

27. The apparatus of claim 22, wherein the means for decomposing the queuing network into separate queuing systems includes means for decomposing the queuing network into decomposed models for each class in a hierarchical manner.

28. The apparatus of claim 27, wherein a decomposed model for class k is based on a decomposed model of classes 1 through k-1.

29. A computer program product in a computer readable medium for allocating resources of a computing system to hosting of a data network site to thereby maximize generated profit, comprising:

first instructions for calculating a total profit for processing requests received by the computing system for the data network site based on at least one service level agreement; and

second instructions for allocating resources of the computing system to maximize the total profit.

30. The computer program product of claim 29, wherein the first instructions include instructions for determining whether processing of each request generates a profit or a penalty for each request received by the computing system for the data network site, wherein a profit is generated when the allocation of resources is such that the request is processed in accordance with the service level agreement and a penalty is generated when the allocation of resources is such that the request is not processed in accordance with the service level agreement.

31. The computer program product of claim 29, wherein the first instructions include instructions for using a cost model in which profit is gained for each request to the data network site that is processed in accordance with a service level agreement and a penalty is paid for each request to the data network site that is not processed in accordance with the service level agreement.

32. The computer program product of claim 29, wherein the requests are classified into one or more classes of requests and each class of request has a corresponding service level agreement from the at least one service level agreement.

33. The computer program product of claim 29, wherein the second instructions include instructions for determining an optimal traffic assignment for routing requests to thereby maximize the total profit.

34. The computer program product of claim 29, wherein the computing system is a web server farm and wherein the resources are servers of the web server farm.

35. The computer program product of claim 34, further comprising third instructions for determining an optimum resource allocation to maximize the total profit.

36. The computer program product of claim 35, wherein the third instructions include:
instructions for modeling the resource allocation as a queuing network;
instructions for decomposing the queuing network into separate queuing systems; and
instructions for summing cost calculations for each of the separate queuing systems.

37. The computer program product of claim 36, further comprising instructions for optimizing the summed cost calculations to maximize generated profit and thereby determine an optimum resource allocation.

38. The computer program product of claim 29, wherein the second instructions include instructions for determining an optimum traffic assignment and an optimum generalized processor sharing coefficient for a class of requests.

39. The computer program product of claim 29, wherein the second instructions include instructions for optimizing a cost function associated with a class of requests.

40. The computer program product of claim 39, wherein the instructions for optimizing the cost function includes instructions for modeling the optimization as a network flow from a source, through sinks representing sites/classes of request and servers/classes of requests, to a supersink.

41. The computer program product of claim 36, wherein the instructions for decomposing the queuing network into separate queuing systems includes instructions for decomposing the queuing network into decomposed models for each class in a hierarchical manner.

42. The computer program product of claim 41, wherein a decomposed model for class k is based on a decomposed model of classes 1 through k-1.